

VISIT...

LANZAROTE
Caliente.COM

Shadows of the Mind

A Search for the Missing Science of Consciousness

ROGER PENROSE

*Rouse Ball Professor of Mathematics
University of Oxford*

Oxford New York Melbourne
OXFORD UNIVERSITY PRESS

1994

laws, and I shall be fairly specific about what the nature of this change must be and of how it might apply to the biology of our brains. Even with our limited present understanding of the nature of this missing ingredient, we can begin to point to where it must be making its mark, and how it should be providing one vital contribution to whatever it is that underlies our conscious feelings and actions.

Though, of necessity, some of the arguments I shall give are not altogether simple, I have tried to make my case as clearly as I can, using only elementary notions where possible. In places, some mathematical technicalities are introduced, but only when they are necessary, or are otherwise helpful for improving the clarity of the discussion. Whilst I have learnt not to expect everyone to be persuaded by the kinds of argument that I shall be presenting, I would suggest, nevertheless, that these arguments do deserve careful and dispassionate consideration; for they provide a case which should not be ignored.

A scientific world-view which does not profoundly come to terms with the problem of conscious minds can have no serious pretensions of completeness. Consciousness is part of our universe, so any physical theory which makes no proper place for it falls fundamentally short of providing a genuine description of the world. I would maintain that there is yet no physical, biological, or computational theory that comes very close to explaining our consciousness and consequent intelligence; but that should not deter us from striving to search for one. It is with such aspirations in mind that the arguments of this book are presented. Perhaps someday the fully appropriate collection of ideas will be brought about. If so, our philosophical outlook can hardly be other than profoundly altered. Yet, all scientific knowledge is a two-edged sword. What we actually *do* with our scientific knowledge is another matter. Let us try to see where our views of science and the mind may be taking us.

1.2 Can robots save this troubled world?

As we open our newspapers or watch our television screens, we seem to be continually assaulted by the fruits of Mankind's stupidity. Countries, or parts of countries, are set against one another in confrontations that may, from time to time, flare into hideous warfare. Excessive religious fervour, or nationalism, or separate ethnic interests, or mere linguistic or cultural differences, or the self-seeking interests of particular demagogues, may result in continuing unrest and violence, sometimes boiling over to outbursts of unspeakable atrocity. Oppressively authoritarian regimes still subjugate their peoples, keeping them in check by the use of death squads and torture. Yet, those who are oppressed, and who might seem to have a common purpose, are often locked in conflict with one another, and when given a freedom that they may

have been long denied, may seem to choose to use that freedom in horribly self-destructive ways. Even in those fortunate countries where there is prosperity, peace, and democratic freedom, resources and manpower are squandered in apparently senseless ways. Is this not a clear indication of the general stupidity of Man? Though we believe ourselves to represent the pinnacle of intelligence in the animal kingdom, this intelligence seems sadly inadequate to handle many of the problems that our own society continues to confront us with.

Yet, the positive achievements of our intelligence cannot be denied. Among these achievements are our impressive science and technology. Indeed, whilst it must be admitted that some of the fruits of this technology are of distinctly questionable long-term (or short-term) value, as is borne witness by numerous environmental problems and a genuine fear of a technology-induced global catastrophe, it is this same technology that has given us our modern society, with its comforts, its considerable freedoms from fear, disease, and need, and with its vast opportunities for intellectual and aesthetic expansion, and for mind-broadening global communication. If this technology has opened up so many potentialities and, in a sense, increased the scope and the power of our individual physical selves, can we not expect much more in the future?

Our senses have been vastly extended by our technology, both ancient and modern. Our sight has been aided and enormously increased in power by spectacles, mirrors, telescopes, microscopes of all kinds, and by video-cameras, television, and the like. Our hearing has been aided, originally by ear-trumpets, but now by tiny electronic devices, and greatly extended by telephones, radio communication, and satellites. We have bicycles, trains, motor cars, ships, and aeroplanes to aid and transcend our natural forms of locomotion. Our memories are helped by printed books, films—and by the huge storage capacities of *electronic computers*. Our calculational tasks, whether simple and routine, or of a massive or sophisticated kind, are also vastly extended by the capabilities of modern computers. Thus, not only does our technology provide us with an enormous expansion of the scope of our *physical* selves, but it also expands our *mental* capabilities by greatly improving upon our abilities to perform many routine tasks. What about mental tasks that are not routine—tasks that require genuine *intelligence*? It is natural to ask whether these also will be aided by our computer-driven technology.

There is little doubt in my own mind that there is indeed, implicit in our (frequently computer-driven) technological society, at least one direction with an enormous potential for enhancing intelligence. I refer, here, to the educational possibilities of our society, which could gain great benefit from different aspects of technology—but only if it is used with sensitivity and understanding. Technology provides the potential, by use of well-produced books, film, television, and interactive computer-controlled systems of various kinds. These, and other developments, provide many opportunities for expanding our minds—or else for deadening them. The human mind is

capable of vastly more than it is often given the chance to achieve. Sadly, these opportunities are all too frequently squandered, and the minds of neither young nor old are provided the openings that they undoubtedly deserve.

But many readers will ask: is there not a rather different possibility for the vast expansion of a mental capability, namely that alien electronic 'intelligence' which is just beginning to emerge from the extraordinary advances in computer technology? Indeed, we already frequently turn to computers for intellectual assistance. There are many circumstances in which unaided human intelligence is far from adequate to assess the probable consequences of alternative actions. Such consequences may lie considerably beyond the scope of human computational powers; thus it is to be expected that the computers of the future will greatly expand this role, where hard computational fact provides an invaluable aid to human intelligence.

Yet may not computers eventually achieve very much more than just this? Many experts claim that computers offer us, at least in principle, the potential for an *artificial* intelligence that will ultimately exceed our own.¹ When computationally controlled robots reach the level of 'human equivalence', then it will not take long, they argue, before they race enormously beyond our own puny level. Only *then*, these experts would claim, shall we have an authority with sufficient intelligence, wisdom, and understanding to be able to resolve the troubles of this world that humanity has created.

How long will it be before this happy state of affairs is to come about? There is no clear consensus among these experts. Some would measure the timescale in terms of many centuries, whilst others claim that this human equivalence is just decades away.² The latter would point to the very rapid 'exponential' growth of computer power and base their estimates upon comparisons between the speed and accuracy of transistors, and the relative slowness and sloppy action of neurons. Indeed, electronic circuits are already over a million times faster than the firing of the neurons in the brain (the rate being some $10^9/s$ for transistors and only about $10^3/s$ for neurons*), and they have an immense precision in timing and accuracy of action that is in no way shared by neurons. Moreover, there is a great deal of randomness in the brain's 'wiring' that, it would seem, could be vastly improved upon by the deliberate and precise organization of electronic printed circuits.

There are some areas where the neuron structure of the brain does provide a numerical advantage over present-day computers, although these advantages might be relatively short lived. It is argued that in total neuron number (some hundreds of thousands of millions), human brains are at the moment ahead of computers with respect to their transistor counts. Moreover, on the average, there are a good deal more *connections* between different neurons than there

*The Intel Pentium chip has over three million transistors on a 'slice of silicon' about the size of a thumbnail, each capable of performing 113 million full instructions per second.

are connections between transistors in a computer. In particular, Purkinje cells in the cerebellum can have up to 80 000 synaptic endings (junctions between neurons), whereas for a computer, the corresponding number is only about three or four at most. (I shall have some comments to make about the cerebellum later; cf. §1.14, §8.6.) Moreover, most of the transistors of today's computers are concerned just with memory and not directly with computational action, whereas it might be the case that with the brain such computational action could be more widespread.

These temporary advantages for the brain could easily be overcome in the future, particularly when massively 'parallel' computational systems become more developed. It is to a computer's advantage that different units can be combined together to form larger and larger ones, so the total number of transistors could, in principle, be increased almost without limit. In addition, there are technological revolutions waiting in the wings, such as the replacing of the wires and transistors of our present computers by appropriate optical (laser) devices, perhaps achieving, thereby, enormous increases in speed, power, and miniaturization. More fundamentally, our brains would appear to be *stuck* with the numbers that we have at present, and we have many further constraints, such as having to grow from a single cell. Computers, on the other hand, can be deliberately constructed so as to achieve all that is eventually needed. Though I shall later be pointing to some important factors that are not yet being taken into account by these considerations (most particularly, a significant level of activity that underlies that of neurons), an impressive-looking case can indeed be made that on any issue of merely computing power, if computers do not have the advantage over brains already, then they *will* certainly have it before too long.

Thus, if we are to believe the strongest of the claims of the most outspoken of the proponents of artificial intelligence, and accept that computers and computer-guided robots will eventually—and even perhaps before too long—exceed all human capabilities, then the computers will be able to do immeasurably more than merely assist *our* intelligences. They will actually have immense intelligences of their own. We could *then* turn to these superior intelligences for advice and authority in all matters of concern—and the humanity-induced troubles of the world could at last be resolved!

But there appears to be another logical consequence of these potential developments that may well strike us as genuinely alarming. Would not these computers eventually make human beings themselves superfluous? If the computer-guided robots turn out to be our superiors in every respect, then will they not find that they can run the world better without the need of us at all? Humanity itself will then have become obsolete. Perhaps, if we are lucky, they might keep us as pets, as Edward Fredkin once said; or if we are clever, we might be able to transfer the 'patterns of information' that are 'ourselves' into robot form, as Hans Moravec (1988) has insisted; or perhaps we will *not* be that lucky and will just *not* be that clever . . .

1.3 The *A*, *B*, *C*, *D*, of computation and conscious thinking

But are the relevant issues merely those of computing power, or of speed, accuracy, or memory, or perhaps of the detailed way in which things happen to be 'wired up'? Might we, on the other hand, be doing something with our brains that cannot be described in computational terms at all? How do our feelings of conscious awareness—of happiness, pain, love, aesthetic sensibility, will, understanding, etc.—fit into such a computational picture? Will the computers of the future actually have *minds*? Does the presence of a conscious mind actually influence behaviour in any way? Does it make sense to talk about such things in scientific terms at all; or is science in no way competent to address issues that relate to the human consciousness?

It seems to me that there are at least four different viewpoints³—or extremes of viewpoint—that one may reasonably hold on the matter:

- A*. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.
- B*. Awareness is a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.
- C*. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
- D*. Awareness cannot be explained by physical, computational, or any other scientific terms.

The point of view expressed in *D*, which negates the physicalist position altogether and regards the mind as something that is entirely inexplicable in scientific terms, is the viewpoint of the mystic; and at least some ingredient of *D* seems to be involved in the acceptance of religious doctrine. My own position is that questions of mind, though they lie very uncomfortably with present-day scientific understanding, should not be regarded as being forever outside the realms of science. If science is yet incapable of saying much that is of significance concerning matters of the mind, then eventually science must enlarge its scope so as to accommodate such matters, and perhaps even modify its very procedures. Whereas I reject mysticism in its negation of scientific criteria for the furtherance of knowledge, I believe that within an expanded science and mathematics there will be found sufficient mystery ultimately to accommodate even the mystery of mind. I shall expand on some of these ideas later on in this book, but for the moment it will be sufficient to say that I am rejecting *D*; and I am attempting to move forward along the path that science has set out for us. If you are a reader who believes strongly that *D*, in some form, must be right, I ask that you bear with me and see how far we can get

along the scientific road—and try to perceive where I believe that this road must ultimately be taking us.

Let us consider what seems to be the other extreme: the viewpoint *A*. Those who adhere to the standpoint that is often referred to as *strong AI* (strong Artificial Intelligence) or sometimes *hard AI*, or *functionalism*,⁴ would come under this heading—although some people might use the term 'functionalism' in a way that could include certain versions of *C* also. *A* is regarded by some as the only viewpoint that an entirely scientific attitude allows. Others would take *A* to be an absurdity that is barely worth serious attention. There are undoubtedly many different versions of viewpoint *A*. (See Sloman (1992) for a long list of alternative computational viewpoints.) Some of these might differ with regard to what kind of thing would be counted as a 'computation' or as 'carrying out' a computation. Indeed, there are also adherents of *A* who would deny that they are 'strong AI supporters' at all, because they claim to take a different view as to the interpretation of the term 'computation' from that of conventional AI (cf. Edelman 1992). I shall address these issues a little more fully in §1.4. For the moment it will be sufficient to take these terms simply to mean the kind of thing that ordinary general-purpose computers are capable of doing. Other proponents of *A* might differ as to how they interpret the meaning of the words 'awareness' or 'consciousness'. Some would not even allow that there is such a phenomenon as 'conscious awareness' at all, whereas others would accept the existence of this phenomenon, but regard it as just some kind of 'emergent property' (cf. also §4.3 and §4.4) that comes along whenever a sufficient degree of complication (or sophistication, or self-reference, or whatever) is involved in the computation that is being performed. I shall indicate my own interpretation of the terms 'consciousness' and 'awareness' in §1.12. Just for now, any differences in possible interpretation will not be greatly important for our considerations.

The strong-AI viewpoint *A* is what my arguments in ENM were most specifically directed against. The length of that book alone should make it clear that, while I do not myself believe that *A* is correct, I *do* regard it as a serious possibility that is worthy of considerable attention. *A* is an implication of a highly operational attitude to science, where, also, the physical world is taken to operate entirely computationally. In one extreme of this view, the universe itself is taken to be, in effect, a gigantic computer;⁵ and appropriate subcomputations that this computer performs will evoke the feelings of 'awareness' that constitute our conscious minds.

I suppose that this viewpoint—that physical systems are to be regarded as merely computational entities—stems partly from the powerful and increasing role that computational simulations play in modern twentieth-century science, and also partly from a belief that physical objects are themselves merely 'patterns of information', in some sense, that are subject to computational mathematical laws. Most of the material of our bodies and brains, after all, is being continuously replaced, and it is just its *pattern* that

persists. Moreover, matter itself seems to have merely a transient existence since it can be converted from one form into another. Even the *mass* of a material body, which provides a precise physical measure of the quantity of matter that the body contains, can in appropriate circumstances be converted into pure energy (according to Einstein's famous $E=mc^2$)—so even material substance seems to be able to convert itself into something with a mere theoretical mathematical actuality. Furthermore, quantum theory seems to tell us that material particles are merely 'waves' of information. (We shall examine these issues more thoroughly in Part II.) Thus, matter itself is nebulous and transient; and it is not at all unreasonable to suppose that the persistence of 'self' might have more to do with the preservation of *patterns* than of actual material particles.

Even if we do not think that it is appropriate to regard the universe as simply being a computer, we may feel ourselves operationally driven to viewpoint $\mathcal{A}$. Suppose that we have a robot that is controlled by a computer and which responds to questioning exactly as a human would. We ask it how it feels, and find that it answers in a way that is entirely consistent with its actually possessing feelings. It tells us that it is aware, that it is happy or sad, that it can perceive the colour red, and that it worries about questions of 'mind' and 'self'. It may even give expression to a puzzlement about whether or not it should accept that *other* beings (especially human beings) are to be regarded as possessing a consciousness similar to the one that it claims to feel itself. Why should we disbelieve *its* claims to be aware, to wonder, to be joyful, or to feel pain, when it might seem that we have as little to go on with respect to other human beings whom we *do* accept as being conscious? The operational argument does, it seems to me, have some considerable force, even if it is not entirely conclusive. If all the *external* manifestations of a conscious brain, including responses to continual questioning, can indeed be completely imitated by a system entirely under computational control, then there would indeed be a plausible case for accepting that its *internal* manifestations—consciousness itself—should be also considered to be present in association with such a simulation.

The acceptance of this kind of argument, which basically is what is referred to as a *Turing test*,⁶ is in essence what distinguishes $\mathcal{A}$ from $\mathcal{B}$. According to $\mathcal{A}$, any computer-controlled robot which, after sustained questioning, convincingly behaves *as though* it possesses consciousness, must be considered *actually* to be conscious—whereas according to $\mathcal{B}$, a robot could perfectly well behave exactly as a conscious person might behave without itself actually possessing any of this mental quality. Both $\mathcal{A}$ and $\mathcal{B}$ would allow that a computer-controlled robot could convincingly *behave* as a conscious person does, but viewpoint $\mathcal{C}$, on the other hand, would not even admit that a fully effective simulation of a conscious person could ever be achieved merely by a computer-controlled robot. Thus, according to $\mathcal{C}$, the robot's actual lack of consciousness ought ultimately to reveal itself, after a sufficiently long

interrogation. Indeed, $\mathcal{C}$ is much more of an *operational* viewpoint than is $\mathcal{B}$ —and it is more like $\mathcal{A}$ than $\mathcal{B}$ in this particular respect.

What about $\mathcal{B}$ then? I think that it is perhaps the viewpoint that many would regard as 'scientific common sense'. It is sometimes referred to as *weak* (or *soft*) AI. Like $\mathcal{A}$, it affirms a view that all the physical objects of this world must behave according to a science that, in principle, allows that they can be computationally simulated. On the other hand, it strongly denies the operational claim that a thing that behaves externally as a conscious being must necessarily be conscious itself. As the philosopher John Searle has stressed,⁷ a computational simulation of a physical process is a very different thing from the actual process itself. (A computer simulation of a hurricane, for example, is certainly no hurricane!) On view $\mathcal{B}$, the presence or absence of consciousness would depend very much upon what actual physical object is 'doing the thinking', and upon what particular physical actions that object is performing. It would be a secondary matter to consider the particular computations that might happen to be involved in these actions. Thus, the action of a biological brain might evoke consciousness, whilst its accurate electronic simulation might well not. It is not necessary, in viewpoint $\mathcal{B}$, for this distinction to be between biology and physics. But the actual *material* constitution of the object in question (say, a brain), and not just its computational action, is regarded as all-important.

The viewpoint $\mathcal{C}$ is the one which I believe myself to be closest to the truth. It is more of an operational viewpoint than $\mathcal{B}$ since it asserts that there are external manifestations of conscious objects (say, brains) that differ from the external manifestations of a computer: the outward effects of consciousness cannot be properly simulated computationally. I shall be giving my reasons for this belief in due course. Since $\mathcal{C}$, like $\mathcal{B}$, goes along with the physicalist standpoint that minds arise as manifestations of the behaviour of certain physical objects (brains—although not necessarily only brains), it follows that an implication of $\mathcal{C}$ is that *not* all physical action can be properly simulated computationally.

Does present-day physics allow for the possibility of an action that is in principle impossible to simulate on a computer? The answer is not completely clear to me, if we are asking for a mathematically rigorous statement. Rather less is known than one would like, in the way of precise mathematical theorems, on this issue.⁸ However, my own strong opinion is that such non-computational action would have to be found in an area of physics that lies *outside* the presently known physical laws. Later on in this book, I shall reiterate some of the powerful reasons, coming from within physics itself, for believing that a new understanding is indeed needed, in an area that lies intermediate between the 'small-scale' level, where quantum laws hold sway, and the 'everyday' level of classical physics. However, it is not by any means universally accepted, among present-day physicists, that such a new physical theory is required.

Thus there are at least two very different standpoints that could come under the heading of $\mathcal{C}$. Some $\mathcal{C}$ -believers would contend that our present physical understanding is perfectly adequate, and that we should look to subtle types of behaviour within conventional theory that might be able to take us outside the scope of what can be achieved entirely computationally (e.g. as we shall examine later: chaotic behaviour (§1.7), subtleties of continuous as opposed to discrete action (§1.8), quantum randomness). On the other hand, there are those who would argue that the physics of today really offers us no reasonable scope for non-computability of the type needed. Later in this book I shall give what I believe are powerful reasons for adopting $\mathcal{C}$ according to this stronger standpoint—which requires some fundamentally new physics to be involved.

Some people have tried to contend that this really places me in camp $\mathcal{D}$, since I am arguing that we must look beyond the reaches of known science if we are ever to find any kind of explanation of the phenomenon of consciousness. But there is an essential difference between this strong version of $\mathcal{C}$ and the viewpoint $\mathcal{D}$ —particularly with regard to the issue of *methodology*. According to $\mathcal{C}$, the problem of conscious awareness is indeed a scientific one, even if the appropriate science may not yet be at hand. I strongly support this viewpoint; I believe that it must indeed be by the methods of science—albeit appropriately extended in ways that we can perhaps only barely glimpse at present—that we must seek our answers. That is the key difference between $\mathcal{C}$ and $\mathcal{D}$, whatever similarities there may seem to be in the corresponding opinions as to what *present-day* science is capable of achieving.

The viewpoints $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, as defined above, are intended to represent extremes, or polarities, of possible stances that one might choose to take. I can accept that some people may feel that their own viewpoints do not fit clearly into any of these categories, but perhaps lie somewhere between them, or cut across some of them. There are certainly many possible gradations of belief between $\mathcal{A}$ and $\mathcal{B}$, for example (see Sloman 1992). There is even a view, not uncommonly expressed, that might best be regarded as a combination of $\mathcal{A}$ and $\mathcal{D}$ (or perhaps $\mathcal{B}$ and $\mathcal{D}$)—a possibility that will actually feature significantly in our later deliberations. According to this view, the brain's action is indeed that of a computer, but it is a computer of such wonderful complexity that its imitation is beyond the wit of man and science, being necessarily a divine creation of God—the 'best programmer in the business'!⁹

1.4 Physicalism vs. mentalism

I should make a brief remark about the use of the words 'physicalist' and 'mentalistic' that are often used to describe opposing standpoints in connection with the issues addressed by $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$. Since $\mathcal{D}$ represents a total denial of physicalism, believers in $\mathcal{D}$ would certainly have to be counted as mentalists. However, it is not at all clear to me where the line between

physicalism and mentalism is to be drawn in relation to the other three viewpoints $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. I think that holders of viewpoint $\mathcal{A}$ would normally be thought of as physicalists, and I am sure that the vast majority of *them* would say so. However, there is something of a paradox lurking here. According to $\mathcal{A}$, the *material* construction of a thinking device is regarded as irrelevant. It is simply the computation that it performs that determines all its mental attributes. Computations themselves are pieces of abstract mathematics, divorced from any association with particular material bodies. Thus, according to $\mathcal{A}$, mental attributes are themselves things with no particular association with physical objects, so the term 'physicalist' might seem a little inappropriate. Viewpoints $\mathcal{B}$ and $\mathcal{C}$, on the other hand, demand that the actual physical constitution of an object must indeed be playing a vital role in determining whether or not there is genuine mentality present in association with it. Accordingly, it might well be argued that these, rather than $\mathcal{A}$, represent the possible physicalist standpoints. However, it seems that such terminology would be at variance with some common usage, the term 'mentalistic' being often regarded as more appropriate for $\mathcal{B}$ and $\mathcal{C}$, since here mental qualities are regarded as being 'real things' and not just as 'epiphenomena' that might arise incidentally when (certain types of) computations are performed. In view of such confusions, I shall tend to avoid the use of the terms 'physicalist' and 'mentalistic' in the discussions that follow, and refer, instead, to the specific viewpoints $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ as defined above.

1.5 Computation: top-down and bottom-up procedures

I have not been at all explicit, thus far, about what I am taking the term 'computation' to mean, in the definitions of $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, of §1.3. What is a computation? In short, one can simply understand that term to denote the activity of an ordinary general-purpose computer. To be more precise, we must take this in a suitably idealized sense: a *computation* is the action of a *Turing machine*.

But what is a Turing machine? It is, indeed, a mathematically idealized computer (the theoretical forerunner of the modern general-purpose computer)—idealized so that it never makes any mistakes and can run on for as long as is necessary, and so that it has an unlimited storage space. I shall be a little more explicit about how Turing machines may be precisely specified in §2.1 and Appendix A (p. 117). (For a much more thorough introduction, the interested reader is referred to the descriptions given in ENM, Chapter 2, or else to Kleene (1952) or Davis (1978), for example.)

The term 'algorithm' is frequently used to describe the action of a Turing machine. I am taking 'algorithm' to be completely synonymous with 'computation' here. This needs a little clarification, because some people take

activity of the cerebellum appears to be entirely unconscious, that it involves perhaps up to one half as many neurons as the cerebrum. Moreover, the Purkinje cells, referred to in §1.2, that have up to 80 000 synaptic connections, are neurons that are found in the cerebellum, so the total number of connections between neurons may well be no fewer in the cerebellum than in the cerebrum. If it is to be the sheer complication of the network of neurons that is regarded as the essential prerequisite for consciousness, then one must ask why consciousness seems to be entirely absent in the actions of the cerebellum. (I shall have some comments to make on this issue later, in §8.6).

Of course, the problems for viewpoint $\mathcal{A}$ referred to in this section have their analogues also for $\mathcal{B}$ and $\mathcal{C}$. On any scientific viewpoint, one would need eventually to address the issue of what it is that underlies the phenomenon of consciousness, and of how qualia can come about. In the later sections of Part II, I shall be attempting to move tentatively towards an understanding of consciousness from the point of view of $\mathcal{C}$.

1.15 Do limitations of present-day AI provide a case for $\mathcal{C}$?

But why $\mathcal{C}$? What evidence is there which can be interpreted as providing direct support for $\mathcal{C}$? Is $\mathcal{C}$ really a serious alternative to $\mathcal{A}$ or $\mathcal{B}$, or even to $\mathcal{D}$? We must try to see what it is that we can actually do with our brains (or minds) when conscious deliberations come into play—and I shall be trying to convince the reader that (sometimes at least) what we do with our conscious thinking is very different from anything that can be achieved computationally. Adherents of $\mathcal{A}$ would be likely to maintain that ‘computing’, in one form or another, is the only possibility—and, as far as the effects on external behaviour are concerned, so also would the adherents of $\mathcal{B}$. On the other hand, adherents of $\mathcal{D}$ might well agree with $\mathcal{C}$ that conscious actions must be things beyond computation, but they would deny the possibility of an explanation of consciousness in any kind of scientific terms. Thus, in order to give support to $\mathcal{C}$, one must try to find examples of mental activity that lie beyond any form of computation, and also try to see how such activity might result from appropriate physical processes. The remainder of Part I will be directed towards the former goal, whilst in Part II, I shall present my attempts to come to terms with the latter.

What kind of mental activity might there be which could be shown to lie beyond computation? As a possible route to this, we could try to examine the present state of artificial intelligence, and try to see what computationally controlled systems are good at and what they are bad at. Of course, the present state of the art of AI may not give a clear indication of what might ultimately be achieved in principle. Even in 50 years, say, things could well be very different from what they are at present. The rapid development of computers

and their applications—only within the *past* 50 years—has been extraordinary. We must certainly be prepared for enormous advances in the future—advances that might possibly come upon us very swiftly indeed. I shall be primarily concerned, in this book, not with the speed of such advances, but with certain fundamental limitations of *principle* that they are subject to. These limitations would apply no matter how many centuries into the future we may be prepared to project our speculations. Thus, we should base our arguments on general principles and not allow ourselves to be unduly influenced by what has been achieved to date. Nevertheless, there could well be clues contained in the successes and failures of the artificial intelligence of today, despite the fact that, so far, there is very little of what could be called a genuinely convincing artificial intelligence—as even the strongest proponents of AI would be prepared to admit.

The main failures of artificial intelligence to date, perhaps rather surprisingly, are not so much in areas where the power of human intellect can itself be extremely impressive—such as where particular human experts can dumbfound the rest of us with their specialist knowledge or their ability to make judgements based on deeply complicated computational procedures—but in the ‘common-sense’ activities that the humblest among us indulge in for most of our waking lives. As yet, no computer-controlled robot could begin to compete with even a young child in performing some of the simplest of everyday activities: such as recognizing that a coloured crayon lying on the floor at the other end of the room is what is needed to complete a drawing, walking across to collect that crayon, and then putting it to its use. For that matter, even the capabilities of an ant, in performing its everyday activities, would far surpass what can be achieved by the most sophisticated of today’s computer control systems. Yet, on the other hand, the development of powerful chess computers provides a striking example in which computers *can* be enormously effective. Chess is undoubtedly an activity where the power of the human intellect is particularly manifest—though exploited to excellence in this way by but a few. Yet chess computer systems now play the game extraordinarily well, and can consistently beat most human players. Even the very best of human experts are now being hard pressed, and may not for long retain what superiority they still possess over the best of chess-playing computers.²² There are also several other areas of expertise, in which computers can compete successfully, or partially successfully, with human experts. Moreover, there are some, such as straightforward numerical computation, in which the capabilities of computers far outstrip the capabilities of humans.

In all these situations, however, it would be hard to maintain that the computer attains any genuine *understanding* of what it is actually doing. In the case of a top-down organization, the reason that the system successfully works at all is not that *it* understands anything, but that human programmers’ understandings (or else the understandings of those human experts upon

whom the programmers depend) have been used in the construction of the program. For a bottom-up organization, it is not clear that there need be any specific understanding whatever, as a feature of the system's actions, on the part either of the device itself or of its programmers—beyond those human understandings that would have gone into the designing of the details of the specific performance-improving algorithms that are involved, and in the very conception that a system can improve its performance with experience whenever an appropriate feedback system is incorporated. Of course, it is not always clear what the term 'understanding' actually means, so some people might claim that on *their* terms, these computer systems actually do possess some kind of 'understanding'.

But is this reasonable? To illustrate a lack of any real understanding by present-day computers, it is interesting to provide, as an example, the chess position given in Fig. 1.7 (taken from an article by Jane Seymore and David

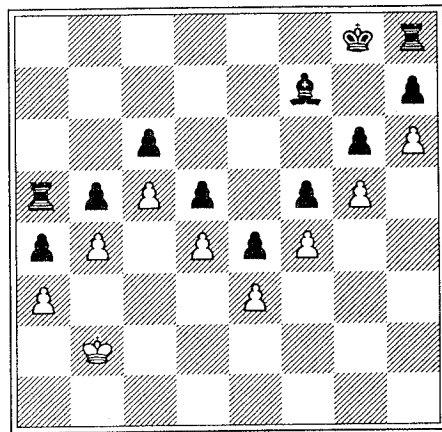


Fig. 1.7. White to play and draw—easy for humans, but Deep Thought took the rook!

Norwood (1993)). In this position, black has an enormous material advantage, to the extent of two rooks and a bishop. However, it is easy for white to avoid defeat, by simply moving his* king around on his side of the board. The wall of pawns is impregnable to the black pieces, so there is no danger to white from the black rooks or bishop. This much is obvious to any

*White is, of course, not necessarily male. See Notes to the reader on p. xvi.

human player with a reasonable familiarity with the rules of chess. However, when the position, with white to move, was presented to 'Deep Thought'—the most powerful chess computer of its day, with a number of victories over human chess grandmasters to its credit—it immediately blundered into taking the black rook with its pawn, opening up the barrier of pawns to achieve a hopelessly lost position!

How could such a wonderfully effective chess player make such an obviously stupid move? The answer is that all Deep Thought had been programmed to do, in addition to having been provided with a considerable amount of 'book knowledge', would be to calculate move after move after move—to some considerable depth—and to try to improve its material situation. At no stage can it have had any actual understanding of what a pawn barrier might achieve—nor, indeed, could it ever have any genuine understanding whatsoever of anything at all that it does.

To anyone with sufficient appreciation of the general way in which Deep Thought or other chess-playing computer systems are constructed, it is no real surprise that it would fail on positions such as that of Fig. 1.7. Not only can we understand something about chess that Deep Thought did not, but we can also understand something of the (top-down) procedures according to which Deep Thought has been constructed; so we can actually appreciate why it should make such a blunder—as well as understanding why it could play chess so effectively in most other circumstances. However, we may ask: is it possible that Deep Thought, or any other AI system, could *eventually* achieve any of the kind of real understandings that we can have—of chess, or of anything else? Some AI proponents might argue that in order for an AI system to gain any 'actual' understanding, it would need to be programmed in a way that involves *bottom-up* procedures in a much more basic way than is usual for chess-playing computers. Accordingly, its 'understandings' would develop gradually by its building up a wealth of 'experience', rather than having specific top-down algorithmic rules built into it. Top-down rules that are simple enough for us to appreciate easily could not, by themselves, provide a computational basis for actual understanding—for we can use our very understandings of these rules to realize their fundamental limitations.

This point will be made more explicit in the arguments given in Chapters 2 and 3. But what about these bottom-up computational procedures? Is it possible that *they* could form the basis of understanding? In Chapter 3, I shall be arguing otherwise. For the moment, we may simply take note of the fact that present-day computer systems do not in any way substitute for genuine human understanding—in any significant area of intellectual expertise where genuine and continuing human understanding and insight seem to be important. This much, I feel sure, would be broadly accepted today. For the most part, the very optimistic early claims²³ that had sometimes been made by proponents of artificial intelligence and promoters of expert systems have not yet been fulfilled.

But these are still very early days, if we are to consider what artificial intelligence might ultimately achieve. Proponents of AI (either *A* or *B*) would maintain that it is just a matter of time, and perhaps some further significant developments in their craft, before important elements of understanding will indeed begin to become apparent in the behaviour of their computer-controlled systems. Later, I shall try to argue in precise terms against this, and that there are fundamental limitations to any purely computational system, whether top-down or bottom-up. Although it might well be possible for a sufficiently cleverly constructed such system to preserve an illusion, for some considerable time (as with Deep Thought), that it possesses some understanding, I shall maintain that a computer system's actual lack of general understanding should—in principle, at least—eventually reveal itself.

For my precise arguments I shall need to turn to some mathematics, the intention being to show that *mathematical* understanding is something that cannot be reduced to computation. Some AI proponents might find this surprising, for they have argued²⁴ that the things that came late in human evolution, like the performing of arithmetical or algebraic calculation, are the things that come most easily to computers, and where computers already outstrip by far the abilities of calculating human beings; whereas those skills that were evolved early, like walking or the interpretation of complicated visual scenes, are things that we perform effortlessly, whilst present-day computers struggle to achieve their unimpressively limited performances. I shall argue very differently. Any complicated activity, which may be mathematical calculations, or playing a game of chess, or commonplace actions—if they have been understood in terms of clear-cut computational rules—are the things that modern computers are good at; but the very understanding that underlies these computational rules is something that is itself beyond computation.

1.16 The argument from Gödel's theorem

How can we be sure that such understandings are not themselves things that can be reduced to computational rules? I shall shortly be giving (in Chapters 2 and 3) some very strong reasons for believing that effects of (certain kinds of) understanding cannot be properly simulated in any kind of computational terms—neither with a top-down, nor a bottom-up organization, nor with any combination of the two. Thus, the human faculty of being able to 'understand' is something that must be achieved by some non-computational activity of the brain or mind. The reader may be reminded (cf. §1.5, §1.9) that the term 'non-computational' here refers to something beyond any kind of effective simulation by means of any computer based on the logical principles that

underlie all the electronic or mechanical calculating devices of today. On the other hand, 'non-computational activity' does *not* imply something beyond the powers of science and mathematics. But it *does* imply that viewpoints *A* and *B* cannot explain how we actually perform all those tasks that are the results of conscious mental activity.

It is certainly a *logical* possibility that the conscious brain (or conscious mind) might act according to such non-computational laws (cf. §1.9). But is it *true*? The argument I shall present in the next chapter (§2.5) provides what I believe to be a very clear-cut argument for a non-computational ingredient in our conscious thinking. This depends upon a simple form of the famous and powerful theorem of mathematical logic, due to the great Czech-born logician Kurt Gödel. I shall need only a very simplified form of this argument, requiring only very little mathematics (where I also borrow from an important later idea due to Alan Turing). Any reasonably dedicated reader should find no great difficulty in following it. However, Gödel-type arguments, used in this kind of way, have sometimes been vigorously disputed.²⁵ Consequently, some readers might have gained an impression that this argument from Gödel's theorem has been fully refuted. I should make it clear this is *not* so. It is true that many counter-arguments have been put forward over the years. Many of these were aimed at a pioneering earlier argument—in favour of mentalism and opposed to physicalism—that had been advanced by the Oxford philosopher John Lucas (1961). Lucas had argued from the Gödel theorem that mental faculties must indeed lie beyond what can be achieved computationally. (Others, such as Nagel and Newman (1958), had previously argued in a similar vein.) My own argument, though following similar lines, is presented somewhat differently from that of Lucas—and not necessarily as support for mentalism. I believe that my form of presentation is better able to withstand the different criticisms that have been raised against the Lucas argument, and to show up their various inadequacies.

In due course (in Chapters 2 and 3), I shall be addressing, in detail, *all* the different counter-arguments that have come to my attention. I hope that my discussion there will serve to correct not only some apparently widespread misconceptions about the significance of the Gödel argument, but also the evidently inadequate brevity of my discussion in ENM. I shall demonstrate that a good many of these counter-arguments are based merely on misconceptions; the remaining ones, which put forward genuine viewpoints that need to be considered in detail, perhaps provide just possible let-outs, in accordance with *A* or *B*, but I shall argue that they nevertheless do *not* really provide *plausible* explanations of what our ability to 'understand' actually allows us to achieve, and that these let-outs would in any case be of little value to AI. Anyone who maintains that all the external manifestations of conscious thought processes *can* be properly computationally simulated, in accordance with either viewpoint *A* or *B*, must find some way of coming to terms, in full detail, with the arguments that I shall give.

ascertaining the truth of statements that apply to *all* natural numbers on the basis of a single computation. In essence, it enables us to deduce that a proposition $P(n)$, that depends on a particular natural number n (such as ‘the sum of the first n hexagonal numbers is n^3 ’), holds for *every* n , provided that we can show, first, that it holds for $n=0$ (or, here, $n=1$) and that we can also show that the truth of $P(n)$ *implies* the truth of $P(n+1)$. I shall not bother the reader with the details of how one would prove that **(E)** never stops, using mathematical induction, but the interested reader might like to try this as an exercise.

Are clear-cut rules, like the principle of mathematical induction, always sufficient to establish the non-stopping nature of computations that in fact do not stop? The answer, surprisingly, is ‘no’. This is one of the implications of Gödel’s theorem, as we shall see shortly, and it will be important that we try to understand it. It is not just mathematical induction that is insufficient. *Any* set of rules *whatever* will be insufficient, if by a ‘set of rules’ we mean some system of formalized procedures for which it is possible to check entirely computationally, in any particular case, whether or not the rules have been correctly applied. This may seem a pessimistic conclusion, for it appears to imply that there are computations that never stop, yet the fact that they never stop cannot ever be rigorously mathematically ascertained. However, this is not at all what Gödel’s theorem actually tells us. What it *does* tell us can be viewed in a much more positive light, namely that the insights that are available to human mathematicians—indeed, to anyone who can think logically with understanding and imagination—lie beyond anything that can be formalized as a set of rules. Rules can sometimes be a partial substitute for understanding, but they can never replace it entirely.

2.5 Families of computations; the Gödel–Turing conclusion $\mathcal{G}$

In order to see how Gödel’s theorem (in the simplified form that I shall give, stimulated also by Turing’s ideas) demonstrates this, we shall need a slight generalization of the kind of statements about computations that I have been considering. Instead of asking whether or not a single computation, such as **(A)**, **(B)**, **(C)**, **(D)**, or **(E)**, ever terminates, we shall need to consider a computation that depends on—or *acts upon*—a *natural number* n . Thus, if we call such a computation $C(n)$, we can think of this as providing us with a *family* of computations, where there is a separate computation for each natural number $0, 1, 2, 3, 4, \dots$, namely the computation $C(0), C(1), C(2), C(3), C(4), \dots$, respectively, and where the way in which the computation depends upon n is itself entirely computational.

In terms of Turing machines, all that this means is that $C(n)$ is the action of some Turing machine on the number n . That is, the number n is fed in on the

machine’s tape as input, and the machine just computes on its own from then on. If you do not feel comfortable with the concept of a ‘Turing machine’, just think of an ordinary general-purpose computer, and regard n as merely providing the ‘data’ for the action of some programmed computer. What we are interested in is whether or not this computer action ever stops, for each choice of n .

In order to clarify what is meant by a computation depending on a natural number n , let us consider two examples:

(F) Find a number that is not the sum of n square numbers

and

(G) Find an odd number that is the sum of n even numbers.

It should be clear from what has been said above that the computation **(F)** will stop *only* when $n=0, 1, 2$, and 3 (finding the numbers $1, 2, 3$, and 7 , respectively, in these cases), and that **(G)** stops for no value of n whatever. If we are actually to ascertain that **(F)** does not stop when n is 4 or larger we require some formidable mathematics (Lagrange’s proof); on the other hand, the fact that **(G)** does not stop for any n is obvious. What are the procedures that are available to mathematicians for ascertaining the non-stopping nature of such computations generally? Are these very procedures things that can be put into a computational form?

Suppose, then, that we have some computational procedure A which, when it terminates,* provides us with a demonstration that a computation such as $C(n)$ actually does not ever stop. We are going to try to imagine that A encapsulates *all* the procedures available to human mathematicians for convincingly demonstrating that computations do not stop. Accordingly, if in any particular case A itself ever comes to an end, this would provide us with a demonstration that the particular computation that it refers to does *not* ever stop. For most of the following argument, it is not necessary that A be viewed as having this particular role. We are just concerned with a bit of mathematical reasoning. But for our ultimate conclusion $\mathcal{G}$, we are indeed trying to imagine that A has this status.

I am certainly not requiring that A can always decide that $C(n)$ does not stop when in fact it does not, but I do insist that A does not ever give us wrong answers, i.e. that if it comes to the conclusion that $C(n)$ does not stop, then in fact it does not. If A does not in fact give us wrong answers, we say that A is *sound*.

It should be noted that if A were actually unsound, then it would be possible

*For the purposes of this argument, I am adopting the viewpoint that if A terminates at all, then this is to signal the achievement of a successful demonstration that $C(n)$ never stops. If A were to ‘get stuck’ for any other reason than ‘success’ in its demonstration, then this would have to qualify as a failure of A to terminate properly. See queries **Q3**, **Q4** below, and also Appendix A (p. 117).

in principle to ascertain this fact by means of some direct calculation—i.e. an unsound A is computationally falsifiable. For if A were to assert erroneously that the computation $C(n)$ does not ever terminate when in fact it does, then the performing of the actual computation $C(n)$ would eventually lead to a refutation of A . (The issue of whether such a computation could ever be performed in practice is a separate matter: it will be discussed under Q8.)

In order for A to apply to computations generally, we shall need a way of coding all the different computations $C(n)$ so that A can use this coding for its action. All the possible different computations C can in fact be listed, say as

$$C_0, C_1, C_2, C_3, C_4, C_5, \dots$$

and we can refer to C_q as the q th computation. When such a computation is applied to a particular number n , we shall write

$$C_0(n), C_1(n), C_2(n), C_3(n), C_4(n), C_5(n), \dots$$

We can take this ordering as being given, say, as some kind of numerical ordering of computer programs. (To be explicit, we could, if desired, take this ordering as being provided by the Turing-machine numbering given in ENM, so that then the computation $C_q(n)$ is the action of the q th Turing machine T_q acting on n .) One technical thing that is important here is that this listing is *computable*, i.e. there is a single* computation $C_\bullet$ that gives us C_q when it is presented with q , or, more precisely, the computation $C_\bullet$ acts on the pair of numbers q, n (i.e. q followed by n) to give $C_q(n)$.

The procedure A can now be thought of as a particular computation that, when presented with the pair of numbers q, n , tries to ascertain that the computation $C_q(n)$ will never ultimately halt. Thus, when the computation A terminates, we shall have a demonstration that $C_q(n)$ does not halt. Although, as stated earlier, we are shortly going to try to imagine that A might be a formalization of *all* the procedures that are available to human mathematicians for validly deciding that computations never will halt, it is not at all necessary for us to think of A in this way just now. A is just *any sound* set of computational rules for ascertaining that some computations $C_q(n)$ do not ever halt. Being dependent upon the two numbers q and n , the computation that A performs can be written $A(q, n)$, and we have:

(H) If $A(q, n)$ stops, then $C_q(n)$ does not stop.

Now let us consider the particular statements (H) for which q is put equal to n . This may seem an odd thing to do, but it is perfectly legitimate. (This is the first step in the powerful 'diagonal slash', a procedure discovered by the highly original and influential nineteenth-century Danish/Russian/German mathe-

*In fact this is achieved precisely by the action of a universal Turing machine on the pair of numbers q, n ; see Appendix A and ENM, pp. 51–7.

matician Georg Cantor, central to the arguments of both Gödel and Turing.) With q equal to n , we now have:

(I) If $A(n, n)$ stops, then $C_n(n)$ does not stop.

We now notice that $A(n, n)$ depends upon just *one* number n , not two, so it must be one of the computations $C_0, C_1, C_2, C_3, \dots$ (as applied to n), since this was supposed to be a listing of *all* the computations that can be performed on a single natural number n . Let us suppose that it is in fact C_k , so we have:

(J) $A(n, n) = C_k(n)$.

Now examine the particular value $n = k$. (This is the second part of Cantor's diagonal slash!) We have, from (J),

(K) $A(k, k) = C_k(k)$

and, from (I), with $n = k$:

(L) If $A(k, k)$ stops, then $C_k(k)$ does not stop.

Substituting (K) in (L), we find:

(M) If $C_k(k)$ stops, then $C_k(k)$ does not stop.

From this, we must deduce that the computation $C_k(k)$ does *not* in fact stop. (For if it did then it does not, according to (M)!) But $A(k, k)$ cannot stop either, since by (K), it is the *same* as $C_k(k)$. Thus, our procedure A is incapable of ascertaining that this particular computation $C_k(k)$ does not stop even though it does not.

Moreover, if we *know* that A is sound, then we *know* that $C_k(k)$ does not stop. Thus, we know something that A is unable to ascertain. It follows that A cannot encapsulate our understanding.

At this point, the cautious reader might wish to read over the whole argument again, as presented above, just to make sure that I have not indulged in any 'sleight of hand'! Admittedly there is an air of the conjuring trick about the argument, but it is perfectly legitimate, and it only gains in strength the more minutely it is examined. We have found a computation $C_k(k)$ that we know does not stop; yet the given computational procedure A is not powerful enough to ascertain that fact. This is the Gödel(–Turing) theorem in the form that I require. It applies to any computational procedure A whatever for ascertaining that computations do not stop, *so long as we know it to be sound*. We deduce that no knowably sound set of computational rules (such as A) can ever suffice for ascertaining that computations do not stop, since there are some non-stopping computations (such as $C_k(k)$) that must elude these rules. Moreover, since from the knowledge of A and of its soundness, we can actually construct a computation $C_k(k)$ that we can *see* does not ever stop, we deduce that A cannot be a formalization of the procedures available to mathemat-

icians for ascertaining that computations do not stop, no matter what A is. Hence:

$\mathcal{G}$ Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth.

It seems to me that this conclusion is inescapable. However, many people have tried to argue against it—bringing in objections like those summarized in the queries **Q1–Q20** of §2.6 and §2.10 below—and certainly many would argue against the stronger deduction that there must be something fundamentally non-computational in our thought processes. The reader may indeed wonder what on earth mathematical reasoning like this, concerning the abstract nature of computations, can have to say about the workings of the human mind. What, after all, does any of this have to do with the issue of conscious awareness? The answer is that the argument indeed says something very significant about the mental quality of *understanding*—in relation to the general issue of computation—and, as was argued in §1.12, the quality of understanding is something dependent upon conscious awareness. It is true that, for the most part, the foregoing reasoning has been presented as just a piece of mathematics, but there is the essential point that the algorithm A enters the argument at two quite different levels. At the one level, it is being treated as just some algorithm that has certain properties, but at the other, we attempt to regard A as being *actually* ‘the algorithm that we ourselves use’ in coming to believe that a computation will not stop. The argument is *not* simply about computations. It is also about how we use our conscious understanding in order to infer the validity of some mathematical claim—here the non-stopping character of $C_k(k)$. It is the interplay between the two different levels at which the algorithm A is being considered—as a putative instance of conscious activity and as a computation itself—that allows us to arrive at a conclusion expressing a fundamental conflict between such conscious activity and mere computation.

However, there are indeed various possible loopholes and counter-arguments that must be considered. First, in the remainder of this chapter, I shall go very carefully through *all* the relevant counter-arguments against the conclusion $\mathcal{G}$ that have come to my attention—these are the queries **Q1–Q20**, that will be addressed in §2.6 and §2.10, which also include a few additional counter-arguments of my own. Each of these will be answered as carefully as I am able. We shall see that the conclusion $\mathcal{G}$ comes through essentially unscathed. Then, in Chapter 3, I shall consider the implications of $\mathcal{G}$ itself. We shall find that it indeed provides the basis for a very powerful case that conscious mathematical understanding cannot be properly modelled *at all* in computational terms, whether top-down or bottom-up or any combination of the two. Many people might find this to be an alarming conclusion, as it may seem to have left us with nowhere to turn. In Part II of this book I shall take a

more positive line. I shall make what I believe to be a plausible scientific case for my own speculations about the physical processes that might conceivably underlie brain action, such as when we follow through an argument of this kind, and how this might indeed elude any computational description.

2.6 Possible technical objections to $\mathcal{G}$

The reader may feel that the conclusion $\mathcal{G}$ is itself quite a startling one, especially considering the simple nature of the ingredients of the argument whereby it is derived. Before we move on to consider, in Chapter 3, its implications with regard to the possibility of building a computer-controlled, intelligent, mathematics-performing robot, we must examine a number of technical points concerning the deduction of $\mathcal{G}$ very carefully. If you are a reader who is not concerned with such possible technical loopholes and are prepared to accept the conclusion $\mathcal{G}$ —that mathematicians are not using a knowably sound algorithm to ascertain mathematical truth—then you may prefer to skip these arguments (for the moment at least) and pass directly on to Chapter 3; moreover, if you are prepared to accept the stronger conclusion that there can be no algorithmic explanation *at all* for our mathematical or other understandings, then you may prefer to pass directly on to Part II—perhaps pausing only to examine the fantasy dialogue of §3.23 (which summarizes the essential arguments of Chapter 3) and the conclusions of §3.28.

There are several points about the mathematics that tend to worry people about the type of Gödel argument given in §2.5. Let us try to sort these out.

Q1. I have taken A to be just a *single* procedure, whereas we undoubtedly use many different kinds of reasoning in our mathematical arguments. Should we not have allowed for a whole list of possible ‘ A ’s?

In fact, there is no loss of generality in phrasing things in the way that I have done. Any finite list $A_1, A_2, A_3, \dots, A_r$ of algorithmic procedures can always be re-expressed as a single algorithm A , in such a way that A will fail to stop only if *all* the individual algorithms $A_1, \dots, A_r$ fail to stop. (The procedure of A might run roughly as follows: ‘Do the first 10 steps of A_1 ; remember the result; do the first 10 steps of A_2 ; remember the result; do the first 10 steps of A_3 ; remember the result; and so on, up to A_r ; then go back to A_1 and do its second set of 10 steps; remember the result; and so on; then the third set of 10 steps, etc. Stop as soon as any of the A_i stops.’) If, on the other hand, the list of the A were infinite, then in order for it to count as an algorithmic procedure, there would have to be a way of generating this entire set $A_1, A_2, A_3, \dots$ in some algorithmic way. Then we can obtain a single A that will do in place of the entire list in the following way:

‘first 10 steps of A_1 ;